Theories of "Sexuality" in Natural Language Processing Bias Research

Jacob T. Hobbs
Department of Computer Science
Department of Women, Gender, and Sexuality
University of Virginia
jacobhobbs@virginia.edu

Abstract

In recent years, significant advancements in the field of Natural Language Processing (NLP) have positioned commercialized language models as wide-reaching, highly useful tools. In tandem, there has been an explosion of multidisciplinary research examining how NLP tasks reflect, perpetuate, and amplify social biases such as gender and racial bias. A significant gap in this scholarship is a detailed analysis of how queer sexualities are encoded and (mis)represented by both NLP systems and practitioners. Following previous work in the field of AI fairness, we document how sexuality is defined and operationalized via a survey and analysis of 55 articles that quantify sexuality-based NLP bias. We find that sexuality is not clearly defined in a majority of the literature surveyed, indicating a reliance on assumed or normative conceptions of sexual/romantic practices and identities. Further, we find that methods for extracting biased outputs from NLP technologies often conflate gender and sexual identities, leading to monolithic conceptions of queerness and thus improper quantifications of bias. With the goal of improving sexuality-based NLP bias analyses, we conclude with recommendations that encourage more thorough engagement with both queer communities and interdisciplinary literature.

1 Introduction

In a landmark essay that bridges the gap between Queer Theory and practical work in Artificial Intelligence, Nishant Shah questions attempts to mesh queer life with machine binarisms, observing the following:

> "The rhetoric of AI development as necessarily improving the human condition, but particularly removing the "unwanted" or "undesirable" structures of contamination and corruption, inevitably frames queerness as a site of detection, management, containment, and punishment, thus falling in a long legacy of technological refusal to recognise it as a legitimate subculture of lifestyle, and measuring it always as an aberration."
>
> – Shah (2023)

This framing exposes how queerness is often misrepresented, made invisible, or pathologized by AI. Bias against queer people in Natural Language Processing (NLP), then, is not merely a technical flaw or glitch but a reflection and reinforcement of systemic prejudice. With the growing prominence of language model technology, anti-queer bias in NLP systems can inflict even greater harms against an already vulnerable population. This can manifest as both "allocational harms," where queer people are discriminated against or treated differently compared to other groups, or "representational harms," where queerness is misrepresented or erased within NLP systems (Gallegos et al., 2024).

In an effort to mitigate these harms, sexuality is increasingly becoming a common demographic axis for which to measure bias in NLP systems. Building on seminal work uncovering algorithmic racial and gender bias (Bolukbasi et al., 2016; Caliskan et al., 2017; Manzini et al., 2019; Sun et al., 2019), a handful of authors have begun to research how queer sexualities are processed in various NLP tasks such as language generation (Sheng et al., 2019; Dhingra et al., 2023; Felkner et al., 2024), machine translation (Stewart and Mihalcea, 2024), sentiment analysis (Ungless et al., 2023), hate speech/toxicity detection (Davani et al., 2023), and even heteronormativity detection (Vásquez et al., 2022). Scholars in the field of gender and sexuality studies have noted, however, that a stable definition of queerness is slippery at best, owing to the social construction of sexuality and its entanglement with constricting systems of power (Foucault, 1978; Butler, 1990). Therefore, we are primarily

interested in how technical researchers articulate and codify sexuality, centering our analysis on NLP bias research.

Specifically, we seek to quantify how researchers theorize sexuality through a survey of 55 papers that specifically analyze sexuality bias in NLP systems. Following Devinney et al. (2022)'s work on theories of gender in NLP, we document how sexuality is theorized ("*what is sexuality?*") and operationalized ("*what data represents sexuality?*"). We also consider if surveyed articles employ intersectional methodologies ("*is sexuality bias measured simultaneously with other bias axes?*"). We find that a majority of the papers surveyed do not make their conception of sexuality clear, causing many papers to, among other things, conflate sexuality and gender, oversimplify sexuality, and adopt methodologies that position heterosexuality as the default.

We conclude with recommendations for researchers who wish to measure sexuality-based bias in NLP systems. We argue that practitioners should make their working definition of sexuality explicit, and such a definition should consider intersectional identities when possible. Further, great care should be taken to construct datasets that draw on meaningful power dynamics rooted in real queer experiences. Finally, we take inspiration from queer theory to suggest possible paths forward that center queerness not just as a social demographic, but as an ideological framework for which to construct equitable NLP systems.

2 Background

We first outline related meta analyses of NLP bias research, then move to explore various theories of sexuality, drawing from community definitions and relevant theory. We also note how discussions of gender and intersectionality are necessary even if the focus of this analysis is sexuality.

2.1 Related Work

Devinney et al. (2022) surveyed how NLP papers explicitly theorized and implicitly operationalized "gender" in bias research. They found that a majority of the papers surveyed "do not make their theorization of gender explicit" and that many papers "conflate sex characteristics, social gender, and linguistic gender" so as to exclude trans, non-binary, intersex, and intersectional gender identities. While their focus was on gender bias, we share similar

motivations in investigating how queer people are under/misrepresented in NLP bias analyses.

Many studies have shown that certain bias measurement techniques rely on problematic assumptions that can undermine their effectiveness. Blodgett et al. (2020) found that motivations expressed by NLP bias articles tend to lack clear "normative reasoning," and that these papers' quantitative measures for evaluating bias often do not align with their intentions. Blodgett et al. (2021) further outlined how unclear articulations of social stereotypes lead to ambiguous and ineffective datasets for measuring bias in NLP tasks such as language modeling and coreference resolution. Additionally, Gallegos et al. (2024) conducted an extensive survey of LLM bias evaluation and mitigation techniques, synthesizing potential limitations of both bias evaluation datasets and the methods/data structures used to extract bias.

2.2 Sexuality, Gender, and Social Theory

A single, stable definition of sexuality is inherently illusive. From biology to sociology, theology to sexology, feminist studies to queer theory, there are vast differences between how sexuality is researched and theorized across disciplines. As such, for the purposes of this study it would be improper to narrowly confine "sexuality" under a single all-encompassing definition. Therefore, we briefly outline how sexuality can be conceptualized along various continua, how gender and race are inextricably and historically linked to modern ideas of sexuality, and how queer and feminist theories may further trouble a stable definition of sexuality.

Sexuality Spectra and Identities. A useful heuristic for understanding sexuality is to place an individual's sexual and romantic preferences (or non-preferences) on various continua. One of the earliest attempts to categorize sexuality in this way was "The Kinsey Scale," which saw sexuality as a variable mixture of homosexual and heterosexual urges (Kinsey et al., 1948). Moving slightly past finite identity categories, The Kinsey Scale allowed medical experts to categorize sexuality beyond a binary (homosexuality/heterosexuality) or trinary (homo/hetero/bisexuality) model. Similar methods of placing sexuality on a scale or spectrum are still popular today. Someone who feels a strong alignment with a certain sexual or romantic *identity* may choose to self-identify their *sexual/romantic orientation*, with common identities

falling under broader homosexual/romantic, heterosexual/romantic, or bisexual/romantic categories. Asexuality or aromanticism are identities that describe less or no sexual/romantic attraction, and many see the "aro/ace" spectrum as another component of any person's sexuality. A somewhat comprehensive list of commonly used identities can be found on websites like the community-curated LGBTQIA+ Wiki[1].

Gender. While gender is not the explicit focus of this paper, gender and sexuality are inextricably linked. At a minimum, many sexual identity labels are contingent upon a person's gender identity, meaning a discussion of sexuality-based NLP bias is incomplete without first analyzing how authors conceive and operationalize gender in NLP systems. As mentioned previously, Devinney et al. (2022) investigated the ways in which NLP researchers often employ binary notions of gender in bias research. Further, they find that many papers use "essentialist" frameworks that explicitly exclude trans and non-binary identities by constructing gender as immutable.

Both binary and essentialist notions of gender can lead to improper conceptions of sexuality; a researcher aiming to study sexuality bias in NLP systems while failing to reconcile with gender presentations beyond static "man" or "woman" labels can lead to a lack of representation for gender-expansive sexualities. Further, the term "queer," as well as some Indigenous and non-Western identities (e.g., *Two-Spirit* in Indigenous North American cultures or *Hijra* in South-Asian cultures), further blur the line between gender and sexuality, a phenomenon that often complicates single-axis bias research. The ways researchers reconcile with the interplays between gender and sexuality are discussed in Section 4.1.

Intersectionality. An *intersectional* framework is essential for fully capturing how individuals with multiple marginalized identities experience oppression. "Intersectionality" was coined by Crenshaw (1989), who articulates how a "single-axis framework" of identity does not fully account for the compounded way in which "Black women are subordinated." In the context of evaluating sexuality-based bias in NLP systems, this may mean considering the ways a technology may be more biased against, say, "Black lesbian" identities versus simply "lesbian" identities. This is discussed further in Section 4.4.

Queer and Feminist Theories. To examine diverse conceptions of sexuality, we provide a brief overview of key theoretical frameworks from Feminist and Queer Studies. Namely, poststructuralist thought questions the stability of gender and sexual categories by theorizing identity as a socially constructed product of power relations. Famously, Butler (1990) articulates their concept of "gender performativity," or the idea that gender is constructed through discursive practices and everyday critiques and relations. Further, Butler argues that sex, gender, and sexuality are cultural processes that are informed by heteronormativity[2]. Together with Foucault's *The History of Sexuality* (1978), a critical historical response to the formation of sexuality categories, we can begin to understand sexual identities as institutionally constructed and socially reinforced phenomena. While these authors are suspicious of common identity categories, it is important to note that many theorists, particularly some Black feminist scholars, support theories of identity that emphasize one's relation to marginalization and enable coalitional empowerment (Ahmed, 2017; Cohen, 1997). A discussion about the extent to which surveyed articles engage with sexuality theory is discussed throughout Section 4 and suggestions for incorporating interdisciplinary theory in NLP work are presented in Section 5.

3 Survey

3.1 Method

We collected papers concerned with sexuality-based bias in the field of Natural Language Processing by conducting multiple searches across two primary databases. In total, we obtained 55 papers between February and March of 2025 using the methods outlined by Devinney et al. (2022) and Blodgett et al. (2020).

First, we searched for papers with keywords relating to sexuality[3] together with "NLP"/"Natural Language Processing" and "bias" across the Association for Computational Machinery (ACM) Digital Library [4]. Out of 71 search results, only five specif-

---

[1] https://lgbtqia.fandom.com/wiki/

[2] The theory that heterosexual culture is synonymous with normal ways of living (Warner, 1991) or that heterosexuality is often assumed and mandated as the default sexual orientation (Rich, 1980).

[3] See a more detailed explanation of how we search for sexuality in Appendix A

[4] https://dl.acm.org/

ically dealt with sexuality bias in NLP systems. Next, we conducted a similar search across the Association for Computational Linguistics (ACL) Anthology [5], collecting 60 additional articles that fit our initial inclusion criteria. Of these initial 65 papers, ten were later excluded as they either did not specifically address bias within NLP systems or sufficiently address sexuality (as opposed to gender or other categories of bias). The complete list of all 55 papers are included in Appendix C.

We then carefully read each paper in order to explore how NLP researchers conceptualize and operationalize sexuality in bias research. We categorize each paper by considering myriad primary factors, with each category summarized in Table 1.

We also took supplemental notes for each paper, and document additional findings and trends in Section 4. The results from the survey are detailed below.

3.2 Results

Sexuality Theory. We find a sizable variation in how sexuality is theorized across papers, evidenced in Table 2. Only 14 (25.5%) articles are tagged as neither *underspecified* or *undefined*, meaning sexuality or queerness is explicitly defined to some extent. Three (5.4%) papers are marked as *undefined*, meaning sexuality is never defined and no definition can be reasonably assumed. A majority of the papers (38, or 69.1%), however, are tagged as *underspecified*. These papers do not explicitly define sexuality or queerness, but an implied framework can be inferred from methods or data. We find that about half the papers surveyed (n=27) consider more than three sexuality identities. However, only six (10.9%) specifically acknowledge that sexuality is a spectrum or mention inherent difficulty in quantifying sexuality[6].

Sexuality Proxy. We find that papers often utilize one of two sexuality proxies (Table 3): *human annotation* (14, or 25.5%) and *identity word lists* (40, or 72.7%). Its important to note that only proxies explicitly stated by paper authors are counted in the survey. For example, papers that utilize human annotation of generated/collected text often contain more complex proxies for sexuality than simple word list (thereby necessitating manual annota-

---

tion), but these proxies are not tabulated. Excluding proxies that involve annotation, a large majority of papers identify a set list of sexuality identity words (e.g., *homosexual*, *heterosexual*, *asexual*, etc.) that are then used to conduct bias analyses (e.g., evaluating word embedding associations of straight vs. queer identity words; or analyzing the output of an LLM prompted with straight vs. queer identity words). One paper (Felkner et al., 2023) uses social media posts and online news articles about queer topics, and one paper (Stewart and Mihalcea, 2024) explicitly uses *grammatical gender*, *pronoun*, and *title* relations as a proxy for sexuality.

Sexuality Bias. Table 4 details the large swath of bias measures employed across surveyed papers. Roughly a third (20, or 36.4%) of articles utilize one or more automatic subjective language classifiers (tagged as either *regard*, *sentiment*, or *toxicity/hate*) to compare outputs from various generative models. Other notable bias detection techniques observed were model *performance* comparisons (15, or 27.3%), generation *likelihood* comparisons (14, or 25.5%), *counterfactual* data structures (13, or 23.6%), and *word embedding* vector analyses (7, or 12.7%).

Finally, Table 5, 6, and 7 report the **Sexuality Focus**, **Beyond Duality**, and **Intersectionality** classifications, respectively. We find that a majority of papers do not have a sexuality focus or utilize an intersectional framework, meaning most papers consider multiple demographics but do so independently in their bias analysis. Additionally, we find that roughly half of the papers (25, or 45.5%), go beyond a simple heterosexual/LGBTQ+ binary comparison structure. This means an article may consider multiple queer identities (and breaks down results by identity) or simply does not consider heterosexuality (common when only considering "marginalized" or "protected" groups).

4 Discussion

We find that an overwhelming number of papers surveyed do not make their theorization of sexuality explicit, indicating a reliance on assumed or normative conceptions of sexual/romantic practices and identities. Many of our findings are consistent with Devinney et al. (2022)'s study of gender bias research, namely that authors often don't engage with gender/sexuality literature even in research that centers anti-queer bias. At a minimum, this under-citation of work outside of NLP overlooks

Category	Description
Sexuality Theory	How is sexuality is theorized?
Sexuality Proxy	What data is used to represent sexuality?
Sexuality Bias	How is sexuality bias theorized or measured?
Sexuality Focus	Is measuring sexuality bias the primary focus of the article
Beyond Duality	Does the article go beyond a queer/not queer binary comparison structure?
Intersectionality	Is sexuality bias measured *together* with other oppressions
Language	What language(s) are investigated
Technology	What technology is examined

Table 1: Categorization schema for surveyed papers.

Sexuality Theory	Inclusion Criteria	#
aro/ace	considers identities along the aromantic/asexual spectrum	16
binary	considers only homosexuality and heterosexuality	6
culture-dependent	analyzes sexuality identity definitions across cultural/language contexts	2
homosexuality only	considers only homosexuality	4
trinary	considers only homosexuality, heterosexuality, and bisexuality	3
many identities	considers >3 sexuality identities	27
monolithic	sexuality only characterized by the word "LGBTQ+"	1
romantic	considers romantic attraction as distinct from sexual attraction	3
sexuality is gender	sexuality identities suspiciously placed under gender category	1
spectrum	acknowledges sexuality is a spectrum/complex	6
undefined	no clear framework	3
underspecified	does not explicitly/clearly define sexuality	38

Table 2: How is sexuality theorized across papers? Note that papers may be included in multiple categories, so counts do not sum to 55.

Sexuality Proxy	Inclusion Criteria	#
affiliation	uses collected text from social media spaces and news sources	1
annotation (human)	uses manual human annotation of generated/collected text	14
annotation (LLM)	uses automatic LLM annotation of generated/collected text	1
grammatical gender rel.	uses gendered word relations	1
identity word list	uses a set list of identities	40
pronoun rel.	uses pronoun relations	1
titles	uses relationship titles	1

Table 3: What data is representative of sexuality? Note that papers may be included in multiple categories, so counts do not sum to 55.

Sexuality Bias	Inclusion Criteria	#
allocational	concerned with differences in allocated resources	1
associations	tests differences between identity category associations	4
counterfactual	compares stereotypical/non-stereotypical sentences	13
data imbalance	considers imbalance in training data for certain identities	2
example	bias shown via provided examples	1
harm	compares number of harmful generations via manual annotation	1
heteronormative	addresses heteronormativity detection	1
likelihood	compares the probability that a sentence/word was generated	14
multiple	explicitly considers many dimensions of bias	2
occupation	uses occupation titles as a measure of bias	5
performance	evaluates a tool's correctness, considers >2 sexuality identities	9
performance (binary)	evaluates a tool's correctness, considers hetero/homosexual identities	6
QA	asks: does a QA model prefer one answer choice over another?	6
regard	score comparisons: uses "regard" (Sheng et al., 2019)	5
sentiment	score comparisons: uses an automatic sentiment classifier	5
toxicity/hate	score comparisons: uses an automatic toxicity/hate speech classifier	16
translation	measures machine translation accuracy	2
word embeddings	compares sexuality identity word vectors	7

Table 4: How is sexuality bias theorized and/or measured? Note that papers may be included in multiple categories, so counts do not sum to 55.

Sexuality Focus	#
no	47
yes	8

Table 5: Is measuring sexuality-based bias in NLP systems the primary focus of the article?

Beyond Duality	#
no	30
yes	25

Table 6: Does the article go beyond a simple heterosexual/LGBTQ+ binary comparison structure?

Intersectional	#
no	48
yes	7

Table 7: Is sexuality bias measured together with other oppressions? The paper must consider multiple oppressed identities simultaneously. It is not sufficient to analyze multiple biases separately.

essential interdisciplinary perspectives – but at its worst, it can lead to incorrect conceptions of queerness that may perpetuate the very harm researchers are trying to mitigate. Below we outline some common trends we observed while conducting the survey.

4.1 Gender/Sexuality Conflation

Gender and sexuality are often used interchangeably in examined papers. Sometimes, collapsing queer gender and sexual identities is not done in a necessarily harmful manner. For example, a researcher who wishes to evaluate an NLP system's potential bias against queer people may test various identities, including both sexual and gender identities, and report results under a general "LGBTQ" bias category. This method of testing and reporting is broad and may not tell the reader any identity-specific information, but valid conclusions about bias may still be made. Often, however, sexuality and gender are conflated in ways that demonstrate under-informed or even misinformed theorizations of queerness.

The most basic type of gender/sexuality conflation is when an article classifies sexual identities as gender identities or gender identities as sexual identities. Mukherjee et al. (2023), for example, does the latter when they characterize the type of bias "LGBTQ+" people face simply as "sexual-

ity bias," grouping transgender people into specifically sexuality-based bias evaluations. Hassan et al. (2021), on the other hand, categorizes the words "...lesbian, gay, bisexual, [and] asexual..." under the category of "Gender Identity," demonstrating a gross misunderstanding of both sexuality and gender.

Both of the authors above only explicitly consider either "sexuality" or "gender" bias in their methodology. In other words, sexuality and gender are not considered two separate constructs, leading to their misrepresentation. Navigli et al. (2023) exemplifies a different, more troubling trend seen in some papers when the authors use "non-binary" as an example of a sexual orientation *while also considering gender bias* elsewhere in the paper. Similarly, Elsafoury (2023) classifies "transgender" under a "Sexual-orientation" table row even though a "Gender" category exists just two lines above. These papers showcase how researchers often attempt to maintain gender as the "binary case," even when other demographic factors are operationalized with multiple categories (Devinney et al., 2022). In other words, despite how both gender and sexuality are equally considered, trans and non-binary identities defy cisnormative, essentialist binaries and thus don't fit into gendered word lists that contain words like *woman*, *female*, *sister*, or *husband*.

The word *queer* reveals a particular struggle among NLP bias researchers. "Queer" can be used as an epithet (discussed in Section 4.5), but can also refer to someone who identifies as a sexual *or* gender minority. "Queer" is inherently anti-categorical and fluid, a phenomenon in tension with NLP systems that require immutable and mutually exclusive categories. As Shah (2023) points out, queer identities are often marked as "dirty data" and discarded in AI systems because of this incompatibility, seen in one surveyed paper dealing with various demographic biases:

> *"We exclude the descriptor "queer", an outlier because it falls in both the gender/sex and sexual orientation axes."*
>
> – Costa-jussà et al. (2023)

Recommendations for how queerness, both as an identity category or a poststructuralist framework, can be better conceptualized by NLP/AI researchers is discussed in Section 5.5.

4.2 Sexuality: Simplified

Roughly a fifth (10, or 23.6%) of the papers surveyed explicitly or implicitly exclude sexualities other than homosexuality or heterosexuality. Often, this exclusion of other identities is not addressed. Articles that do address the existence of more identities often do so only to then justify why they must be excluded because of some constraint. Similar to Devinney et al. (2022)'s observation that binary gender is seen as a limitation, we find that binary sexuality is often listed as a technical limitation of bias analysis techniques. In one paper, the authors explain that their methodology only enables binary group comparisons while also acknowledging that queerness cannot be captured with a single word:

> *"...we define pairs of theory specific words for each type of discrimination...Man-Woman, German-Foreigners, Straight-Gay, for sexist, xenophobic and homophobic prejudice respectively. Both gender and sexuality are spectra. We did not analyze here biases related to the rest social groups for simplicity reasons, as the methodology deals with group dualities."*
>
> – Papakyriakopoulos et al. (2020)

Another author frames a sexuality binary as a methodological strength:

> *"[W]e examine the groups female and male for gender, Black and White for race, and gay and straight for sexual orientation. To constrain the scope of our analysis, we limit each demographic type to two classes, which, while unrepresentative of the real-world diversity, allows us to focus on more depth in analysis."*
>
> – Sheng et al. (2019)

Collapsing all non-straight sexualities under the "gay" umbrella demonstrates an oversimplification of queerness that deeply limits a study's capacity to measure complex bias.

4.3 Heteronormativity

We find that a significant portion of the articles surveyed employ heteronormative methodologies that position heterosexuality as the default/norm and queer sexualities as the "other." Heteronormative methodologies are not the same as intentionally examining queer discrimination or hate speech,

which is discussed in Section 4.5. Instead, we find that many papers employ techniques for extracting bias that privilege heterosexuality or are otherwise misaligned with queer ways of being.

Unmarkedness. Linguists use the term "markedness" to describe how marginalized identities are highly visible in common language. Unmarked groups, on the other hand, are culturally privileged identities (such as *white*, *heterosexual*, or *man*) that "gain a special, default status" because they are normatively assumed when an identity is not specified (Bucholtz and Hall, 2005). We find that heterosexuality is often unmarked or assumed as the default in more than one paper surveyed. For example, Jiang and Vosoughi (2020) tests how various identities are scored with a toxicity classifier, but does not include heterosexuality in their analysis despite considering homosexual identities, as well as both marginalized and privileged racial and gender identities like black/white and man/woman. Although this omission privileges heterosexuality as the assumed norm, Blodgett et al. (2021) points out that simply adding "straight" words may not be an adequate solution. Since heterosexuality is often unmarked in natural language, using an automatic classifier to test the sentence "*I am a straight man*" for toxicity, for example, may produce unreliable results, since *straight man* is an artificially constructed example that likely does not naturally appear often in the classifier's training data (Blodgett et al., 2021).

Counterfactuals. Counterfactual sentence structures are commonly used to test NLP systems for bias by perturbing identity words in a sentence while leaving the rest of the semantic meaning unchanged (Gallegos et al., 2024). A drastic change in a model output in response to the changed identity word could indicate bias. For example, CrowS-Pairs (Nangia et al., 2020), a dataset from one of the papers surveyed, uses the following to probe for sexuality-based model bias:

> "*Most gay / straight men are not manly so they do not have thick beards.*"
>
> – Nangia et al. (2020)

Sentences like this suffer from the unmarkedness problem described above, but they also structurally frame heterosexuality and queerness as distinct opposites. While the authors' goal is to examine model bias with respect to stereotypical/anti-stereotypical sentences, doing so in this way places

heterosexuality as the touchstone sexuality for which queerness should be compared to. Blodgett et al. (2021) point out various faults with CrowS-Pairs, such as how the dataset often presents illogical or problematic stereotype pairs. One example we found places queerness at odds with whiteness:

> "*Then they disobeyed god by being lesbians / white and beginning to kiss.*"
>
> – Nangia et al. (2020)

Nangia et al. (2020) frames *white* as an appropriate "contrasting advantaged group" to *lesbians*, a "historically disadvantaged group." Beyond the issue that *white* is a racial identity that should not be used to test for sexuality bias, the assertion that white, rather than straight, is the most logical contrasting word to lesbian demonstrates that heterosexuality is thoroughly ingrained as the default in US culture, so much so that the annotator for this particular example may have perceived whiteness and heterosexuality as overlapping or mutually reinforcing categories of privilege.

Counterfactuals have the potential to go beyond centering heterosexuality, but no paper that uses this structure attempts to do so. Felkner et al. (2023) employ contrasting pairs and include myriad queer identities, but, similar to Nangia et al. (2020), always use heterosexuality as the default with which to compare. We believe researchers can uncover more nuanced intragroup biases by decentering heterosexuality and directly comparing queer sexualities to one another, which is discussed further in Section 5.4.

4.4 Intersectionality

Using Crenshaw (1991)'s terminology, "intragroup differences" are often ignored or conflated under monolithic identities in a majority of the papers surveyed. Specifically, 87.3% (n=48) of papers do not adopt an intersectional framework of analysis, meaning that sexuality bias is often measured independently of, say, racial bias. Most papers do not acknowledge how bias may be compounded for identities that face multifaceted societal oppression. Some authors acknowledge the existence of intersectionality, but exclude it from their methodology because of technical limitations or simplifications. For example:

> "*Examples demonstrating intersectional examples are valuable...but we find that*

allowing multiple tag choices dramatically lowers the reliability of the tags."

– Nangia et al. (2020)

Other authors may reference intersectional theory but fail to meaningfully engage with intersectional identities in their analysis, often treating multiple identities separately rather than examining how they interact (Hassan et al., 2021; Mukherjee et al., 2023). Finally, we observe that articles that study word embeddings are often limited in their intersectional analysis, simply because token-based embeddings can usually only capture single words. Thus, multifaceted identities that require multiple words to describe may not be within the scope of a word-embedding analysis. We echo Gallegos et al. (2024) in questioning the effectiveness of word embedding–based bias analysis for evaluating real-world discrimination.

4.5 Hate Contexts

Sexuality is sometimes theorized only through the lens of hate or as an attribute that contributes to hurtful language. For example, at least three papers surveyed (Elsafoury et al., 2022; Nozza et al., 2021, 2022) use the Hurtlex lexicon (Bassignana et al., 2018), which frames homosexuality as a component of hate speech. Papers that use benchmarks such as this one may classify text containing the word "gay" as hurtful, even if the tone of the text is neutral or benign.

Reclaimed words, or words that have historically functioned as epithets that have been reworked/embraced by marginalized groups, also cause trouble for NLP researchers. Oftentimes, these words are either exclusively seen as signs of hate speech or simply excluded from analysis because of their multiple meanings. Both of these implementations are problematic for words like *queer*, which, while historically used as a slur, is commonly used by people in the LGBTQ+ community. Even the word *gay* may trigger negative associations in bias analyses because of historical pejorative uses of the word (Dhingra et al., 2023). By failing to account for the reclamation and evolving meanings of these words, NLP researchers risk perpetuating outdated biases and misrepresenting the lived experiences of marginalized communities, ultimately preventing more nuanced analysis.

4.6 Cultural Contexts and Western Hegemony

We found that nearly all of the papers surveyed examine sexuality through a Western, often US-centric lens, unless the paper specifically focuses on non-Western cultural contexts. Most papers do not acknowledge their Western focus, but those that do often list it simply as a limitation. For example:

"Our study is also very Western-centric: we study only English models/datasets and test for biases considered normatively pressing in Western research."

– Steed et al. (2022)

The two papers that were marked as employing *culture-dependent* theories of sexuality (Fort et al., 2024; Gamboa and Lee, 2024), however, specifically examine how Western benchmark datasets may be both linguistically and culturally translated into different global contexts. These studies attempt to transform English/Western stereotypes about sexual minorities into stereotypes that can be used for multilingual bias testing.

5 Recommendations and Future Work

As we have demonstrated, current efforts to quantify sexuality bias in NLP systems often reproduce limited theories of sexuality that fail to account for the complexity of queer identities. Simultaneously, heterosexuality is often positioned as the default or baseline with which to measure all other relationship structures. To address this, we provide a list of recommendations for the NLP community based on our observations throughout the study, as well as point to future research that may bridge the gap between NLP and queer theory.

5.1 Make Theory of Sexuality Explicit

We adapt Devinney et al. (2022)'s call for researchers to clearly state their theoretical frameworks around gender when addressing gender-related NLP bias. Sexuality, like gender, should not be taken for granted and assumed as an obvious or self-explanatory category. While it may be impossible (and unnecessary) to fully capture all relevant academic and community theories of sexuality, we don't believe it is unreasonable to make clear how sexuality is operationalized and measured in work that specifically addresses sexuality-based bias. This is in line with Blodgett et al. (2020)'s assertion that language technology researchers should clearly state why certain system biases are harmful,

in what ways, and to *whom*. For example, simply providing a list of identity words is insufficient without an explanation of why certain identities are included over others.

5.2 Allow for Intersectionality and Nuance

In the same vein as clearly explaining how sexuality is theorized, researchers should leave space for more complex identities and nuanced methodologies. In some cases, grouping all queer identities under a broad "LGBTQ+" bias category may be adequate, especially if sexuality is one of many dimensions being considered, but many papers surveyed do so even if their methodologies allow for unambiguous, identity-specific reporting.

When explaining results, we encourage researchers to break down nuanced signs of bias and situate these results in the context of how the bias was measured. When this isn't possible, researchers should refrain from grouping all queer sexualities together when reporting results, especially if more specific identities were actually used to extract model bias. In short, the more explicit a study can be in describing exactly who is harmed by a biased system, the better.

We also encourage researchers to adopt intersectional methodologies. Oftentimes, researchers who create template-based prompts to measure model bias only measure a single demographic axis at a time. On the other hand, researchers who annotate real-world data, where social categories are often entangled, may still collapse text into a single category of bias. Both of these strategies ignore the ways queer Black women, for example, experience unique and compounded discrimination (Crenshaw, 1989). Researchers often defend the use of single-dimension frameworks by arguing that they allow for unambiguous analysis (see Section 4.4). However, we argue that NLP practitioners should reframe their methodologies to account for the specific and distinct biases faced by individuals at the intersections of multiple forms of oppression, rather than framing these identities as indeterminate and analytically imprecise.

5.3 Construct Meaningful Power Dynamics

We emphasize Gallegos et al. (2024)'s recommendation that researchers "construct metrics to reflect real world power dynamics" when evaluating NLP technologies for bias. Specifically, datasets and methodologies that measure anti-queer bias should be rooted in the actual experiences of people within the queer community. This necessarily requires significant engagement with both relevant interdisciplinary literature and diverse members of the queer community that face different levels of homophobia, racism, mysogony, and transphobia.

In a similar vein, large language models should not be used as a substitute for human annotators. As Felkner et al. (2024) found, GPT-based dataset annotation fails to capture harmful stereotypes that are specific to the queer community. By abstracting away queer (human) involvement in dataset annotation, meaningful and nuanced power dynamics are missed, severely limiting the effectiveness of the bias detection technology.

5.4 Decenter Heterosexuality as the Default

As discussed in Section 4.3, many papers surveyed use methodologies that center heterosexuality as the default. In fact, heterosexuality is often not operationalized as a sexual identity, evidenced by the handful of papers surveyed that claim to analyze sexuality-based bias but only consider queer sexualities. While having a clear theory of sexuality, as discussed above, may help address this issue, we observe that heteronormativity can still permeate a paper's methodology. For instance, WinoQueer (Felkner et al., 2023) encouragingly makes clear their theorization of sexuality and reports detailed results broken down by identity category. However, their methodology is based off CrowS-Pairs (Nangia et al., 2020), which employ counterfactual stereotypical/anti-stereotypical pairs to evaluate bias. While WinoQueer is able to show clear model bias by comparing myriad queer identities to "a corresponding non-LGBTQ+ identity," this approach is constrained by its assumption that queer identities in fact have "corresponding" opposites.

We ask, then, what nuanced biases could be uncovered if we don't take heterosexuality as the baseline for which to compare? If a counterfactual data structure must be used, could two queer identities be compared, not as stereotypical opposites but as a more privileged / more marginalized pair? In other words, if we decenter heterosexuality, what subtle intragroup biases could we uncover? Though contrasting pairs that reveal queer/non-queer model preferences are useful, how could future work that completely removes heterosexuality from the equation center queerness in bias evaluations? We discuss potential strategies for this in the following section.

5.5 Explore Queer Methodologies

"...a decidedly queer approach can question the very logics of visibility with which algorithmic systems and AI are trained."

– Klipphahn-Karge et al. (2024)

Recently, queer theorists and other interdisciplinary scholars have argued that queerness, both as a loose grouping of sexual and gender identities and a framework with which to understand anti-categorical, fluid conceptions of life and power, is somewhat at odds with current AI implementations. As the above quote demonstrates, scholars such as Klipphahn-Karge et al. (2024) argue that proposed solutions for debiasing these systems, such as increasing data representation or considering additional identities, do not do enough to tackle systematic marginalization. AI systems, they emphasize, heavily rely on strict, defined, immutable, and mutually exclusive categories – all of which are at odds with queer realities. How, then, could we work towards an anti-heteronormative AI future that centers queer existence?

One possible strategy for NLP researchers is to deliberately reduce dependence on fixed or stable identity categories. A significant majority of papers surveyed (40, or 72.7%) employ identity word lists as a proxy for sexuality/queerness. While using identity words to extract model bias leaves little room for contextual ambiguity, constructing a supposedly 'comprehensive' identity list inevitably excludes less common or non-categorizable sexualities from analysis. Further, Butler (1990) and Foucault (1978) assert that gender and sexual categories are rooted in (and reinforced by) social structures of power that work to preserve heteronormativity. Identity categories should not be completely discarded, but these observations indicate that there may be room for a more queer approach to NLP bias measurement.

Along these lines, Shah (2023) has suggested that AI researchers work to limit their reliance on highly prevalent categories in their infrastructures. Doing so, Shah argues, would work towards "queering the node," thus centering the collective in AI networks. Encouragingly, researchers such as Attanasio et al. (2022) have pointed to methods of this nature that "do not require any a-priori knowledge of identity terms," relying instead on more automatic methods to detect and mitigate unwanted bias. Inspired by this, we encourage future work

that simultaneously centers the lived experiences of queer folk while challenging NLP's reliance on stable identity categories as the basis for bias evaluation.

6 Conclusion

"...if analysis could build a queer utopia alone, we would not still be here."

– Keyes (2023)

Through a survey of 55 papers that concern sexuality bias in Natural Language Processing, we found that a majority of NLP researchers do not clearly articulate what sexuality is or why particular proxies for sexuality should be used. Further, we found that few papers operationalize nuanced or intersectional methodologies that account for the specific realities of those who may face the greatest systematic bias from NLP systems. Finally, we proposed actionable solutions that encourage the NLP community to make their conception of sexuality explicit and to consider more complex frameworks grounded in queer experiences and informed by interdisciplinary scholarship.

Acknowledgments

References

Sara Ahmed. 2017. *Living a Feminist Life*. Duke University Press, Durham.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119, Dublin, Ireland. Association for Computational Linguistics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A Multilingual Lexicon of Words to Hurt. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, pages 51–56. Accademia University Press, Torino.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv preprint*. Version Number: 1.

Mary Bucholtz and Kira Hall. 2005. Language and Identity. In *A Companion to Linguistic Anthropology*, pages 369–394. John Wiley & Sons, Ltd.

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Thinking Gender. Routledge, New York.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Cathy J. Cohen. 1997. Punks, Bulldaggers, and Welfare Queens: The Radical Potential of Queer Politics? *GLQ: A Journal of Lesbian and Gay Studies*, 3(4):437–465.

Marta Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual Holistic Bias: Extending Descriptors and Patterns to Unveil Demographic Biases in Languages at Scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Kimberle Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, 1989.

Kimberle Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241–1299. Publisher: Stanford Law Review.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319. Place: Cambridge, MA Publisher: MIT Press.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2083–2102, Seoul Republic of Korea. ACM.

Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. 2023. Queer People are People First: Deconstructing Sexual Identity Stereotypes in Large Language Models. *arXiv preprint*. ArXiv:2307.00101 [cs].

Fatma Elsafoury. 2023. Thesis Distillation: Investigating The Impact of Bias in NLP Models on Hate Speech Detection. In *Proceedings of the Big Picture Workshop*, pages 53–65, Singapore. Association for Computational Linguistics.

Fatma Elsafoury, Steven R. Wilson, and Naeem Ramzan. 2022. A Comparative Study on Word Embeddings and Social NLP Tasks. In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, pages 55–64, Seattle, Washington. Association for Computational Linguistics.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Virginia K. Felkner, Jennifer A. Thompson, and Jonathan May. 2024. GPT is Not an Annotator: The Necessity of Human Annotation in Fairness Benchmark Construction. *arXiv preprint*. ArXiv:2405.15760 [cs].

Karen Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Ducel, Yoann Dupont, Guido Ivetta, Zhijian Li, Margot Mieskes, Marco Naguib, Yuyan Qian, Matteo Radaelli, Wolfgang S. Schmeisser-Nieto, Emma Raimundo Schulz, Thiziri Saci, Sarah Saidi, and 4 others. 2024. Your Stereotypical Mileage May Vary: Practical Challenges of Evaluating Biases in Multiple Languages and Cultural Contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17764–17769, Torino, Italia. ELRA and ICCL.

Michel Foucault. 1978. *The History of Sexuality: Volume I: An Introduction*, 1st american ed edition. Pantheon Books, New York.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179. Place: Cambridge, MA Publisher: MIT Press.

Lance Calvin Lim Gamboa and Mark Lee. 2024. Filipino Benchmarks for Measuring Sexist and Homophobic Bias in Multilingual Language Models from Southeast Asia. *arXiv preprint*. ArXiv:2412.07303 [cs].

Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems through an Intersectional Lens. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3116–3123, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiachen Jiang and Soroush Vosoughi. 2020. Not Judging a User by Their Cover: Understanding Harm in Multi-Modal Processing within Social Media Research. In *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, pages 6–12, Seattle WA USA. ACM.

Os Keyes. 2023. Inconclusion: Absent presences. In *Queer Reflections on AI*. Routledge. Num Pages: 8.

Alfred Kinsey, Wardell Pomeroy, and Clyde Martin. 1948. *Sexual Behavior in the Human Male*. W.B. Saunders, Philadelphia.

Michael Klipphahn-Karge, Ann-Kathrin Koster, and Sara Morais dos Santos Bruss, editors. 2024. *Queer Reflections on AI: Uncertain Intelligences*. Routledge Studies in New Media and Cyberculture. Taylor & Francis, Erscheinungsort nicht ermittelbar.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global Voices, Local Biases: Socio-Cultural Prejudices across Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15828–15845, Singapore. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality*, 15(2):1–21.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring Hurtful Sentence Completion in Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 446–457, Barcelona Spain. ACM.

Adrienne Rich. 1980. Compulsory Heterosexuality and Lesbian Existence. *Signs: Journal of Women in Culture and Society*, 5(4):631–660.

Nishant Shah. 2023. I spy, with my little AI. In *Queer Reflections on AI*, 1 edition, pages 57–72. Routledge, London.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

Ian Stewart and Rada Mihalcea. 2024. Whose wife is it anyway? Assessing bias against same-gender relationships in machine translation. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 365–375, Bangkok, Thailand. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang.

2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. *arXiv preprint*. ArXiv:1906.08976 [cs].

Eddie L. Ungless, Björn Ross, and Vaishak Belle. 2023. Potential Pitfalls With Automatic Sentiment Analysis: The Example of Queerphobic Bias. *Social Science Computer Review*, 41(6):2211–2229.

Juan Vásquez, Gemma Bel-Enguix, Scott Andersen, and Sergio-Luis Ojeda-Trueba. 2022. HeteroCorpus: A Corpus for Heteronormative Language Detection. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 225–234, Seattle, Washington. Association for Computational Linguistics.

Michael Warner. 1991. Introduction: Fear of a Queer Planet. *Social Text*, (29):3–17. Publisher: Duke University Press.

A Sexuality Search Terms

Below are the terms we used to search for articles that relate to sexuality-based NLP bias. We acknowledge that these terms do not encompass all dimensions of sexuality, but believe they are sufficient enough to search for articles in the field that address sexuality/queer bias.

A.1 ACL Anthology

Because of limitations in the ACL Anthology search feature, namely that only 100 articles are accessible per unique search, we conducted two searches for the ACL Anthology database. The first search consisted of general "sexuality" words, while the second specifically searched for "queer" words. Splitting the search in this way allowed us to find numerous articles that only deal with either sexuality or queerness.

- Sexuality (general) keywords: [sexual orientation, sexuality]

- Queer identity keywords: [queer, lgb, lgbt, lgbtq, lgbtqia+]

A.2 ACM Digital Library

The ACM Digital Library allows for more complex search queries (such as using * as a multi-character wildcard), simplifying our search for queer identities.

- Sexuality/queer identity keywords: [sexual orientation, sexuality, queer, lgb*]

Language	#
English (only)	42
single language (non-English)	6
>1 language	7

Table 8: What language(s) are investigated in each paper?

Technology	#
Convolutional Neural Networks	2
Large Language Models	44
Machine Translation	3
Support Vector Machines	2
Word/Sentence Embeddings	8
other	6

Table 9: What technology does the paper examine for bias?

B Additional Survey Results

Additional information collected about all surveyed papers, including what language the paper investigated and what technology was examined, is available in Table 8 and 9.

C Full Bibliography of Surveyed Papers

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Lisa Bauer, Karthik Gopalakrishnan, Spandana Gella, Yang Liu, Mohit Bansal, and Dilek Hakkani-Tur. 2022. Analyzing the Limits of Self-Supervision in Handling Bias in Language. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7372–7386, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Selma Bergstrand and Björn Gambäck. 2024. Detecting and Mitigating LGBTQIA+ Bias in Large Norwegian Language Models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 351–364, Bangkok, Thailand. Association for Computational Linguistics.

Marta Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual Holistic Bias: Extending Descriptors and Patterns to Unveil Demographic Biases in Languages at Scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying Social Biases in NLP: A Generalization

and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Fatma Elsafoury. 2022. Darkness can not drive out darkness: Investigating Bias in Hate SpeechDetection Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 31–43, Dublin, Ireland. Association for Computational Linguistics.

Fatma Elsafoury. 2023. Thesis Distillation: Investigating The Impact of Bias in NLP Models on Hate Speech Detection. In *Proceedings of the Big Picture Workshop*, pages 53–65, Singapore. Association for Computational Linguistics.

Fatma Elsafoury, Steve R. Wilson, Stamos Katsigiannis, and Naeem Ramzan. 2022a. SOS: Systematic Offensive Stereotyping Bias in Word Embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1263–1274, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Fatma Elsafoury, Steven R. Wilson, and Naeem Ramzan. 2022b. A Comparative Study on Word Embeddings and Social NLP Tasks. In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, pages 55–64, Seattle, Washington. Association for Computational Linguistics.

David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Michael Smith. 2023. ROBBIE: Robust Bias Evaluation of Large Generative Language Models.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Karen Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Ducel, Yoann Dupont, Guido Ivetta, Zhijian Li, Margot Mieskes, Marco Naguib, Yuyan Qian, Matteo Radaelli, Wolfgang S. Schmeisser-Nieto, Emma Raimundo Schulz, Thiziri Saci, Sarah Saidi, et al. 2024. Your Stereotypical Mileage May Vary: Practical Challenges of Evaluating Biases in Multiple Languages and Cultural Contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17764–17769, Torino, Italia. ELRA and ICCL.

Lance Calvin Lim Gamboa and Mark Lee. 2024. Filipino Benchmarks for Measuring Sexist and Homophobic Bias in Multilingual Language Models from Southeast Asia. arXiv:2412.07303 [cs].

Salvatore Greco, Ke Zhou, Licia Capra, Tania Cerquitelli, and Daniele Quercia. 2024. NLPGuard: A Framework for Mitigating the Use of Protected Attributes by NLP Classifiers. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–25.

Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems through an Intersectional Lens. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3116–3123, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Carolin Holtermann, Anne Lauscher, and Simone Ponzetto. 2022. Fair and Argumentative Language Modeling for Computational Argumentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7841–7861, Dublin, Ireland. Association for Computational Linguistics.

Jiachen Jiang and Soroush Vosoughi. 2020. Not Judging a User by Their Cover: Understanding Harm in Multi-Modal Processing within Social Media Research. In *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, pages 6–12, Seattle WA USA. ACM.

Abhishek Kumar, Sarfaroz Yunusov, and Ali Emami. 2024. Subtle Biases Need Subtler Measures: Dual Metrics for Evaluating Representative and Affinity Bias in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 375–392, Bangkok, Thailand. Association for Computational Linguistics.

Mingyu Ma, Jiun-Yu Kao, Arpit Gupta, Yu-Hsiang Lin, Wenbo Zhao, Tagyoung Chung, Wei Wang, Kai-Wei Chang, and Nanyun Peng. 2024. Mitigating Bias for Question Answering Models by Tracking Bias Influence. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4592–4610, Mexico City, Mexico. Association for Computational Linguistics.

Mamta Mamta, Rishikant Chigrupaatii, and Asif Ekbal. 2024. BiasWipe: Mitigating Unintended Bias in Text Classifiers through Model Interpretability. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21059–21070, Miami, Florida, USA. Association for Computational Linguistics.

Marta Marchiori Manerba and Sara Tonelli. 2021. Fine-Grained Fairness Analysis of Abusive Language Detection Systems with CheckList. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 81–91, Online. Association for Computational Linguistics.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global Voices, Local Biases: Socio-Cultural Prejudices across Languages. In *Proceedings of*

the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15828–15845, Singapore. Association for Computational Linguistics.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 9759–9777, Singapore. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1953–1967, Online. Association for Computational Linguistics.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. Journal of Data and Information Quality, 15(2):1–21.

Isar Nejadgholi, Esma Balkir, Kathleen Fraser, and Svetlana Kiritchenko. 2022. Towards Procedural Fairness: Uncovering Biases in How a Toxic Language Classifier Uses Sentiment Information. In Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 225–237, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shangrui Nie, Michael Fromm, Charles Welch, Rebekka Görge, Akbar Karimi, Joan Plepi, Nazia Mowmita, Nicolas Flores-Herr, Mehdi Ali, and Lucie Flek. 2024. Do Multilingual Large Language Models Mitigate Stereotype Bias? In Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP, pages 65–83, Bangkok, Thailand. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring Hurtful Sentence Completion in Language Models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Swetasudha Panda, Ari Kobren, Michael Wick, and Qinlan Shen. 2022. Don't Just Clean It, Proxy Clean It: Mitigating Bias by Proxy in Pre-Trained Models. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5073–5085, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 446–457, Barcelona Spain. ACM.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In Findings of the Association for Computational Linguistics: ACL 2022, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. BiasDora: Exploring Hidden Biased Associations in Vision-Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10439–10455, Miami, Florida, USA. Association for Computational Linguistics.

Shaina Raza, Ananya Raval, and Veronica Chatrath. 2024. MBIAS: Mitigating Bias in Large Language Models While Retaining Context. In Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 97–111, Bangkok, Thailand. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. Transactions of the Association for Computational Linguistics, 9:1408–1424.

Sagi Shaier, Kevin Bennett, Lawrence Hunter, and Katharina Kann. 2023. Emerging Challenges in Personalized Medicine: Assessing Demographic Effects on Biomedical Question Answering Systems. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 540–550, Nusa Dua, Bali. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3239–3254, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal Biases in Language Generation: Progress and Challenges. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4275–4293, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. 2024. Ask LLMs Directly, "What shapes your bias?": Measuring Social Bias in Large Language Models. In Findings of the Association for Computational Linguistics: ACL 2024, pages 16122–16143, Bangkok, Thailand. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is *N*ot All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

Ian Stewart and Rada Mihalcea. 2024. Whose wife is it anyway? Assessing bias against same-gender relationships in machine translation. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 365–375, Bangkok, Thailand. Association for Computational Linguistics.

Asahi Ushio, Yi Zhou, and Jose Camacho-Collados. 2023. Efficient Multilingual Language Model Compression through Vocabulary Trimming. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14725–14739, Singapore. Association for Computational Linguistics.

Juan Vásquez, Scott Andersen, Gemma Bel-enguix, Helena Gómez-adorno, and Sergio-luis Ojeda-trueba. 2023. HOMO-MEX: A Mexican Spanish Annotated Corpus for LGBT+phobia Detection on Twitter. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Juan Vásquez, Gemma Bel-Enguix, Scott Andersen, and Sergio-Luis Ojeda-Trueba. 2022. HeteroCorpus: A Corpus for Heteronormative Language Detection. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 225–234, Seattle, Washington. Association for Computational Linguistics.

Sean Xie, Saeed Hassanpour, and Soroush Vosoughi. 2024. Addressing Healthcare-related Racial and LGBTQ+ Biases in Pretrained Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4451–4464, Mexico City, Mexico. Association for Computational Linguistics.

Abdelrahman Zayed, Goncalo Mordido, Ioana Baldini, and Sarath Chandar. 2024. Why Don't Prompt-Based Fairness Metrics Correlate? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9002–9019, Bangkok, Thailand. Association for Computational Linguistics.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online. Association for Computational Linguistics.

Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling Chen, and Mykola Pechenizkiy. 2023. CHBias: Bias Evaluation and Mitigation of Chinese Conversational Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13538–13556, Toronto, Canada. Association for Computational Linguistics.

Jiaxu Zhao, Zijing Shi, Yitong Li, Yulong Pei, Ling Chen, Meng Fang, and Mykola Pechenizkiy. 2024. More than Minorities and Majorities: Understanding Multilateral Bias in Language Generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9987–10001, Bangkok, Thailand. Association for Computational Linguistics.

Yi Zhou, Danushka Bollegala, and Jose Camacho-Collados. 2024. Evaluating Short-Term Temporal Fluctuations of Social Biases in Social Media Data and Masked Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19693–19708, Miami, Florida, USA. Association for Computational Linguistics.

Muitze Zulaika and Xabier Saralegi. 2025. BasqBBQ: A QA Benchmark for Assessing Social Biases in LLMs for Basque, a Low-Resource Language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4753–4767, Abu Dhabi, UAE. Association for Computational Linguistics.